ANALYSIS OF THE MEDICARE ECONOMIC INDEX

PROGRESS REPORT 2
OPTION YEAR 2
CONTRACT NO. 500-78-0052

Submitted by:
R. Berry

Project Manager:
R. Budnitz
Director,
Energy and Environmental
Technologies Division

Teknekron Research Inc.
2118 Milvia Street
Berkeley, California 94704

Submitted to:
Dr. Peter McMenamin
Dept. of Health & Human Services
330 Independence Ave., S.W.
Washington, D.C. 20201

STATEMENT OF COST AND PERSONNEL RESPONSIBLE FOR REPORT

This Report is made pursuant to Contract No. 500-78-0052. The charge to the Department of Health and Human Services for the work resulting in this report (inclusive of the amount so charged for any prior reports submitted under this contract) is $50,831. The names of the persons, employed or retained by the contractor with managerial or professional responsibility for such work or for the content of the report, are as follows

1. Robert R. Berry.

INTRODUCTION

This second Progress Report describes work performed through Sept. 15, 1981, for the Second Option Year of Contract No. 500-78-0052. The work required under this Contract consists of the use of the Physician Practice Cost Survey to analyze the variation in physicians' costs, Area 1, and the general characteristics of physicians' practices, Area 4. Teknekron Research, Inc. has already completed the required work for the Base Year and the First Option Year of the Contract through use of the input data from the 1976 and 1977 Surveys. The Second Option Year continues the same type of analysis using the 1978 Survey.

I. COMBINATION OF THE 1976, 1977 AND 1978 SURVEYS

Teknekron has completed its programming efforts of combining the 1976, 1977, and 1978 Surveys. Constructing this joint year file is necessary in order to perform an analysis of the physician firm over time. The first Progress Report briefly reviewed Teknekron's efforts to join the Survey files for the three separate years, and this Progress Report will discuss each step of the file construction in some detail.

Although the basic structure of these annual Survey files has not changed through the years, the definition, location, and coding rules for many of the variables have changed from year to year. In addition, several variables appear in only one or two of the three Surveys. These changes require transforming each annual Survey file into a uniform file whose variables are similarly defined and identically located. These annual files were then concatenated into one all-year file. This concatenated file includes all cases from each of the three Surveys, a total of 13831 observations of which 3482 cases come from the 1976 Survey, 4865 from the 1977 Survey, and 5484 from the 1978 Survey. Editing to eliminate cases with data likely to be erroneous was not performed separately on each of the annual files. Instead, such editing will take place on the concatenated file even though some of the rules for editing for certain variables are time dependent (e.g., costs).

Creating this concatenated file required several steps. Work on each of these steps was performed separately for each year in order to prepare three separate but uniform input files for an SPSS joint year save file. Programs used in the steps prior to the creation of the SPSS save files were written in PL/I primarily to take advantage of its facility for string interpretation and efficient record input/output. The statistical save files used SPSS rather than SAS or BMD because project end-users other than the programmer were more familiar with SPSS and because SAS has a SPSS to SAS conversion procedure which is efficient and inexpensive. The SAS to SPSS conversion program is not publicly supported at the Stanford University Computer Center where the software was developed. First, the specialty code was corrected in those cases where the physician's actual specialty differed from the specialty which NORC had initially assumed. NORC used the A.M.A. specialty designation as reported by the physician to construct a random sample stratified by specialty and region but upon interviewing the physician, NORC discovered that these designations were incorrect in approximately 100 cases in each of the three years. Second, because much of the file construction and analysis is done by specialty, the file was sorted by this corrected specialty code and a sequence number assigned to each case on output. Third, each record was partitioned into three sub-records: one with sequence number, specialty code, and specialty fee and reimbursement data; a second with sequence number, FIPS codes, and AMA-source variables such as specialty board data, school

codes etc.; and a third with sequence number and the entire Survey plus ARF variables which is the remainder and major portion of the record. This partitioning reduced the input/output counts and space requirements in subsequent processing.

Fourth, the first sub-record for the specialty fee/reimbursement was condensed. Each physician was asked about specialty fees/reimbursements only for a pair of procedures appropriate to own specialty but not of course about procedures specific to other specialties. For simplification of documentation, NORC had allocated sufficient space on its output record for all of the specialty fee responses for each case even though each physician answered for only a pair of fees. NORC's approach allowed each variable to be defined uniquely but consumed fifteen times the minimum space. NORC assigned missing values for the fees other than the physician's own. We defined two new sets of varibles, one for each of the two procedures about which the physician provided fee and reimbursement information. These sets, labelled Specialty Procedure One and Specialty Procedure Two and containing the five fee and reimbursement variables, occupy the same record location for each physician but are defined differently depending on the physician's specialty. (In the SPSS save file, separate specialty procedure value labels were coded to identify the procedure.) Statistical checks were made to insure the accuracy of this step by comparing fees and reimbursements across years.

In the fifth step, the second sub-record was transformed and

A.M.A. data added to the record. A categorical school quality variable was created; also, the license year and A.M.A. specialty codes were converted to numeric values. These recodings were checked.

In the second part of this step, the 1976 A.M.A. special count of physicians by county (numbers of specialists in each of four activity levels) was spliced to this small sub-record of transformed variables. The A.M.A. file was transformed prior to this matching operation; the original file was considerably condensed by restricting the number of specialties and the number of activity counts. To splice these two files required matching records on a county basis; the FIPS state and county codes in each of the records was used for the matching. To perform this task, each file was sorted by these FIPS codes. Then, a program searched the A.M.A. file to find a county whose FIPS code matched the FIPS code on the Survey record; this search required a 'look-ahead' check. Less than ten records in any Survey file had FIPS codes which could not be matched. This program carefully checked the matching of the FIPS codes from the Survey and the A.M.A. files, assigning a missing value if no match was possible for a particular record. The program identified the un-matched Survey records and produced a count of the total matched records. Matching only this small end section of each Survey record with a record from the A.M.A file reduced the input/output and space requirements. At this step, a count of physicians by county by specialty was output in order to determine the distribution of physicians sampled. In the final operation in this step, the

matched file was re-sorted by the case sequence number.

In the sixth step, the three sub-records were joined. this operation was straightforward because each of the three input files was identically ordered; again these three files are first the one with AMA physician counts and identifiers; second, the condensed specialty fee file, and third, the remainder of the original record (i.e. the Survey and the ARF), which was in fact the major portion of the NORC-supplied record. This spliced file constitutes the master or base file for each year; it contains all the variables which can be used in any of the statistical analyses. Using this master file as input, a separate program selected variables as output, which were then in turn input to the statistical programs. This program which selects variables from the master file is a very simple PLI record input/output program and is designed to be altered quite readily for selection of a larger or smaller set of variables. The program allows easy re-calculation of the record length and re-definition of both the input and output structure. Those variables currently outputted include all those from the Survey, including both the follow-up and specialty fees, the Survey and AMA-source personal identifiers, and selected ARF variables. The Area Resource File (ARF) is a file maintained by the Health Resource Administration of H.H.S. For each county, the ARF includes data on population, general population characterisitics from the preceding Census, numbers of medical personnel and facilities, health status indicators, Medicare and general health care expenditures, prevailing charge rates under Medicare, etc. Because counties

vary considerably in size, the county charactersitics will not be uniformly related to the individual practice. Practices in smaller and certainly practice in more homogenous communities will probably be more strongly influenced by the county-wide characteristics. Nevertheless, certain ARF variables were selected as proxies for the characteristics of the area in which the practice was located. These variables include county population, SMSA population, counts of physicians and nurses , population characteristics such as median income, years of education,etc., Medicare expenditures and enrollees, hospital and nursing home size and utilization, prevailing charges, etc.

The construction of an SPSS sav file constituted the seventh step. A separate statistical save file program was written for each Survey year. These programs are not identical because the input format and coding rules changed from year to year. Also, where a variable appeared in just one year, it was assigned a missing value for the other years. All variables have mnemonic labels and although the three save files are separate, naming conventions are uniform for variables and value labels in order to reduce the problem of mis-identifying variables when selecting a set for input to the joint year file. All missing values are recoded to a uniform value of -9 for consistency. Recoding and variable transformation statements are grouped logically (not by the order of the input variable's location) in order to simplify future changes. More complete documentation of the SPSS save file differences will appear in the next Progress Report.

II. DETERMINATION OF SAMPLE SIZE FOR THE 1978 SURVEY

Accompanying this Progress Report is a separate volume of tables which show a count of the number of cases which have non-missing values for single variables and groups of variables. The function of these tables is to assist O.R.D.'s preparation of a report describing some of the variables in the 1978 Survey. Primarily this report will present tables and descriptive statistics (e.g. counts, means) for such variables in the Survey as the various types of hours and visits broken down by specialty, by region and by type of practice.

Preparation of this report requires first selecting the variables of interest and then finding a set of cases which has no missing values for the selected variables. Following this approach of using the same sample for each analysis in the report will allow a comparison of the results across the separate tables and statistics. The limitation of this approach is that the number of cases can become quite small if the variables with missing values do not substantially overlap As an extreme example, if those cases with missing values for hours are not those with missing values for fees, then using a single sample of cases which has no missing values for either set of variables could eliminate much of the original sample from the analysis. The larger the number of variables and the greater the problem of non-overlapping missing values, the smaller will be the final set of cases used

for analysis. Only by imputing valid values in cases where a value is missing can this problem be overcome and the sample essentially expanded. However, the impact of the problem can be lessened if the set of variables chosen for the analysis takes into account the problem of missing values. The tables in the accompanying volume will assist this process of variable selection by indicating the number of c ases with missing values for each variable and for many groups of variables in the Survey. These tables are divided into two groups: those showing a breakdown of the single variables by specialty and by region, and those showing a breakdown of groups of variables by specialty. The first group of tables includes most of the variables from the 1978 Survey; it excludes some of the practice type variables and all the procedure variables. Examination of the tables in the Section labelled, 'Set 1, Single Variables by Specialty' reveals that about 5 percent of the cases have missing values for the hours and visit variables. The visit variables have slightly greater numbers of missing values for corresponding categories (e.g. patient hours has only 3.7 percent missing but patient visits has 4.5). As indicated below, those missing are not the same for all these hours and visit variables, leading to a much higher cumulative total.

The cost variables have a much higher percentage of missing values, between 14 and 36 percent for the nine main cost variables where the percent missing are in parentheses: rent (23), equipment (34), supplies (27), employee salaries (14), auto expenses (34), malpractice premiums (16), net income (23), gross

income (29) and employee fringe benefit ratio (36). The greatest number of missing values occurs in auto expenses, cost of equipment, and employee fringe benefits with this last variable having the most missing, 1981. In general, the missing values appear to be fairly evenly distributed across specialties. (Note to determine the percent of missing cases by specialty, use for the total number of cases the count in the breakdown by 'NUMDS' for which there are no missing values.) One cost related variable, the number of examining rooms, has almost no missing values (only 103). The individual employee count, hour and salary variables have at most 17 percent missing (clerk salaries) with most only missing about 5 percent. Note that one important cost variable has numerous missing values, firm-wide malpractice costs (over 70 percent).

The insurance variables have between 12 and 15 percent missing except for the Medicaid variable where the percent exceeds 30; note that for these variables not applicable was recoded as missing and this recode had the greatest impact on that variable. If one choose to recode not applicable as zero, this will add 1134 valid cases to the Medicaid variable and can be done by refering to the previous indicator variable, whether the physician accepts Medicaid patients or not.

The fee and reimbursement variables show a much higher percentage of missing value cases. The greatest number of missing values occurs in the reimbursments for the two specialty fees. About 28 percent of the office fee responses and about 38 percent of the hospital fee responses are missing. Approximately half of the

physicians did not respond to questions on reimbursement from Medicare and Medicaid for these office and hospital follow-up visits. These reimbursement variables are important to include in the analysis but doing so will drastically reduce the sample size.

The second set of tables, labelled 'Set 1, Single Variables by Region', displays the breakdown of the same set of single variables but by the for Census regions. These tables help identity the expected cell size for most of the variables of interest.

The third set of tables presents the breakdowns of groups of variables by specialty. This set of tables is labelled 'Set 2, Groups of Variables by Specialty'. Variables were grouped by type by adding all variables in the group and dividing this sum by itself. Thus, the value of the variable itself is of no interest, but what is of interest is the count of valid cases by specialty. The groups of variables are practice type, hours, visits, minority patients, insurance participation, fees and reimbursements, costs (in foour groups) and the personal background variables. nNeither practice type, hours nor visits has a substantial count of missing values; the count for vists is 572 of the total 5484, which is slightly over ten percent. The insurance participation variables show a much larger missing value count of nearly forty percent. A total of 2006 cases are missing for one of the insurance variables. Most of the missing values are in the Medicaid variable, and the subsequent set of tables presents the insurance participation group without

Medicaid and the count of missing values falls to 1020. The variable labelled, 'Tier 1', groups the practice type, hour, visit, minority and insurance sub-groups. This variable shows a missing value count of 2569 or nearly fifty percent.

The next set of variables shows the missing value count for the fees. The count for the office follow-up fee is 3395, for the hospital follow-up fee is 3697, for the first specialty fee 3595 and for the second specialty fee 3872 Thus, the percent of missing values exceeds sixty percent for the follow-up fees and seventy percent for the second specialty fee. These fee group variables include all the reimbursement variables, too. Combining the fee group variables dramatically reduces the sample size to 508, or a total of 4845 missing values.

The first cost group variable, COSTN1, includes the variables which are part of the Index and gross income and total deductions. This group only has 1879 valid cases; this low number is partially due to the combined effect of missing values for auto expenses, net income and gross income, each of which has well over 1000 missing value cases. Mostly the reduction is due to the variable total deductions which had over 2000 missing values. This group could be re-constructed to eliminate total deductions and auto expenses and this would likely increase the number of cases considerably. The second group of cost variables, labelled COSTN2, includes all the varaibles in the first group plus the total employee count, the physician salaries and hours. This addition reduces the valid case count to 1417 from 1879. Adding the employee cost and hour group to the second

cost group produces COSTN3 and this restricts the number of valid cases to 1417. When the other cost-related variables such as pensions and outside income have been added (through COSTN4), the sample size for valid cases has been reduced to 274 cases.

THE ADDITION OF THE COST GROUP VARIABLES TO THE TIER1 GROUP YIELDS A SAMPLE SIZE OF 1157; NEARLY EIGHTY PERCENT OF THE CASES ARE MISSING. TIER4 INCLUDES ALL THE COST VARIABLES AND THIS FURTHER REDUCES THE SAMPLE SIZE TO 187. TIER5 ADDS THE PERSONAL BACKGROUND VARIABLES AND PRODUCES A COUNT OF 31 VALID CASES.